ANOVA Test

The previous lecture looked at the F-distribution. The F-test is used to determine if population A has a larger variance than population B. A second use of F distributions it to compare the means of 3+ populations. This is called an Analysis of Variance (ANOVA) test.

Knowing whether 3+ populations have the same mean can be useful:

- If you have 3+ bags of parts, such as 3904 transistors, you can tell if the three bags came from the same (or similar) production runs.
- You can tell if the mean of a process is consistent or if it is changing, and
- You can check if different groups of data can be combined into a larger group of if they should be kept separate (due to having different means).

In this lecture, we'll go over

- What the Analysis of Variance test is when you have access to the raw data and
- An approximation for when you only know the mean, standard deviation, and sample size of the data.

We'll then use the ANOVA test to determine if

- Global temperatures had the same mean over the decades of 1880, 1890, and 1900
- If global temperatures changed in the decades of 1880, 1930, 1980, and 2020

ANOVA Test

The basic idea behind an ANOVA test is this:

Assume you have samples from three populations with unknown means and variances

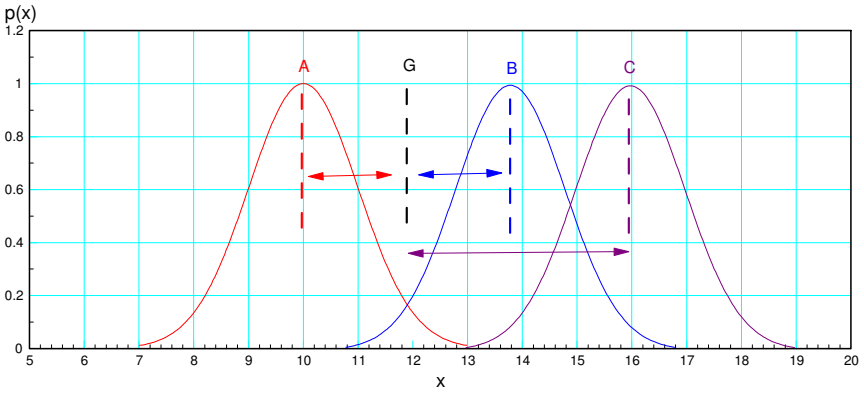
- Each population will have a mean and a variance
- The whole sample size will have a mean and a variance

Now take two measurements:

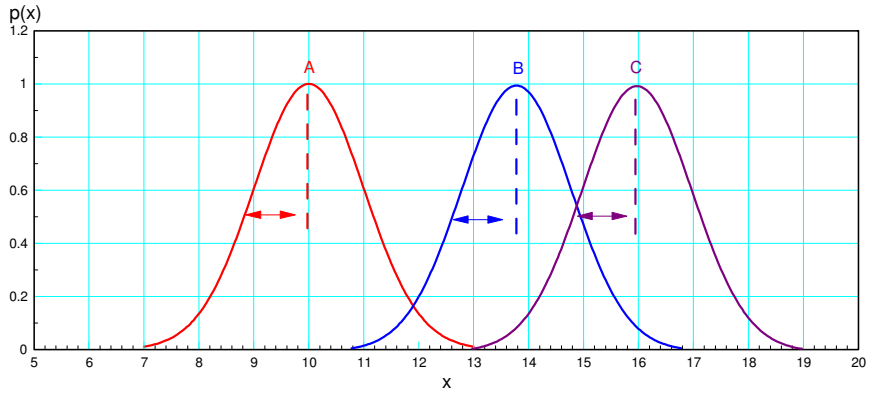
- One measures the mean sum-squared distance from each population to the global mean (mean sum squared between populations, or MSSB)
- And the other measures the mean sum squared distance from each population to that populations' mean (mean sum square within populations, or MSSW)

From these two measurements, form an F-statistic comparing the variances:

$$F = \frac{\text{MSS}_b}{\text{MSS}_w} = \frac{\text{mean sum of squares between data sets}}{\text{mean sum of squares within data sets}}$$



MSSb: The weighted distance (squared) from each populations mean to the global mean (G)



MSSw: The distance (squared) from each data point to it's respective mean

If all populations have the same mean, the two numbers should be the same (and the ratio should be one)

$$F = \frac{MSS_b}{MSS_w} \approx 1$$

If one (or more) populations has a mean which is significantly different, then the ration should be much larger than one:

$$F = \frac{MSS_b}{MSS_w} > 1$$

ANOVA Equations: Variation #1 (nonstandard Method)

Assume you have three populations (A, B, and C) and that you have access to the raw data (meaning you know the actual measurements. Define

k	the number of data sets (assume k = 3 here)
a_i, b_i, c_i	samples from data sets A, B, and C
$\overline{A}, \overline{B}, \overline{C},$	the means of each data set
n_a, n_b, n_c	the number of data points in each data set
s_a^2, s_b^2, s_c^2	the variance of each data set
$N = n_a + n_b + n_c$	the total number of data points
$\overline{G}$	the global average (average of all data points)
s_g^2	the global variance (variance of all data points treated as one population)

MSSb: Mean Sum Squared Distance Between Columns

MSSb measures the distance from the data points to the global mean. If you have access to the raw data, then MSSb is the variance of the entire data set:

$$MSS_b = \left(\frac{1}{N-1}\right)\left(\sum (a_i - G)^2 + \sum (b_i - G)^2 + \sum (c_i - G)^2\right)$$

or equivalently, MSSb is the variance of the entire data set

$$MSS_b = var(\{A, B, C\})$$

with N-1 degrees of freedom

$$dof = N - 1$$

MSSw: Mean Sum Squared Distance Within Columns

MSSw measures the total variance of each population. Two (equivalent) ways to find MSSw are:

$$MSS_w = \left(\frac{1}{N-k}\right)\left(\sum \left(a_i - \overline{A}\right)^2 + \sum \left(b_i - \overline{B}\right)^2 + \sum \left(c_i - \overline{C}\right)^2\right)$$

or

$$MSS_w = \left(\frac{1}{N-k}\right)\left((n_a - 1)s_a^2 + (n_b - 1)s_b^2 + (n_c - 1)s_c^2\right)$$

These are the same equations since the variance of a population is by definition

$$s_a^2 = \frac{1}{n_a - 1}\sum \left(a_i - \overline{A}\right)^2$$

The degrees of freedom for MSSw is (N - k)

$$dof = (n_a - 1) + (n_b - 1) + (n_c - 1)$$
$$= N - k$$

F-value: The F-value is then the ratio

$$F = \frac{MSS_b}{MSS_w}$$

Once you get an F-value, you can convert this to a probability using an F-table (or StatTrek). Larger F-values indicate a higher chance of rejecting the null hypothesis (or putting it another way, the probability that the means of the populations are different).

To get an idea of what the F-statistic looks like as the mean of C varies, consider the following example.

ANOVA Example #1:

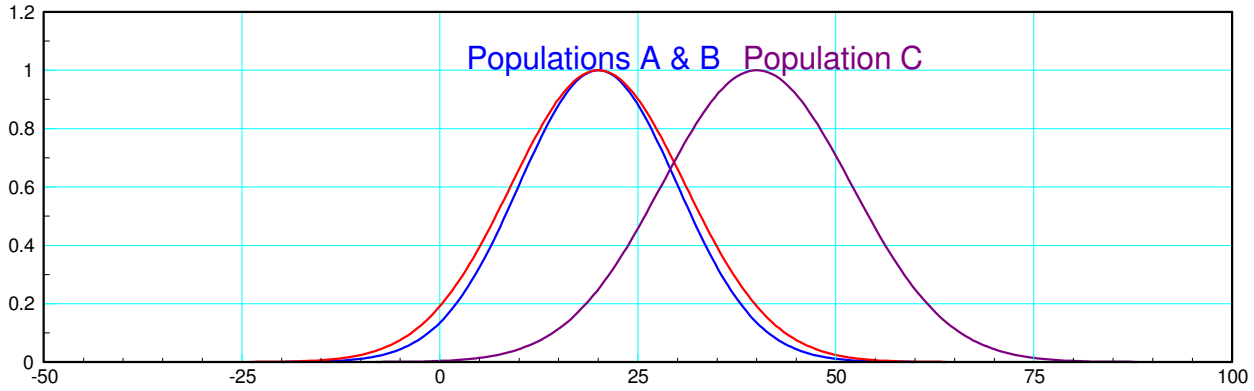
Let's consider three populations where the means are the same:

$$A \sim N(20, 10^2)$$

$$B \sim N(20, 11^2)$$

$$C \sim N(\overline{C}, 12^2)$$

and $\overline{C}$ varies from 20 to 50.



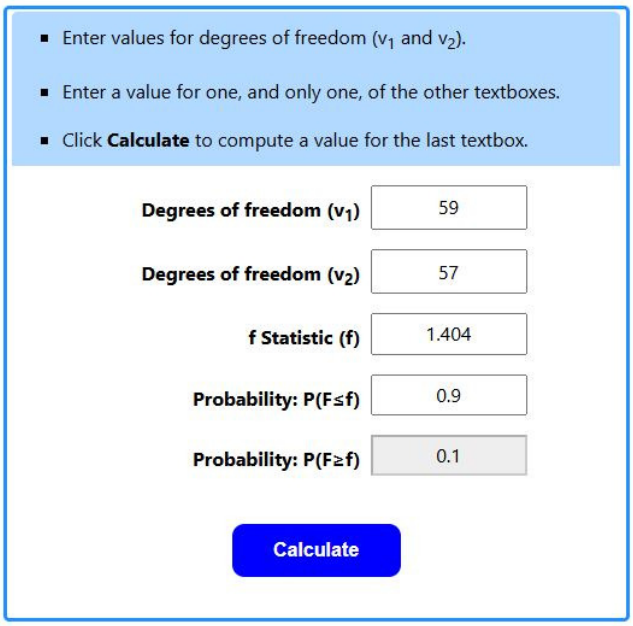
Example 1: pdf for three populations. A and B have a mean of 20. C's mean varies from 20 to 50

If you collect 20 samples from each population, the F-value you're looking for is either

- F > 1.404 for p = 90%
- F > 2.818 for p = 99%

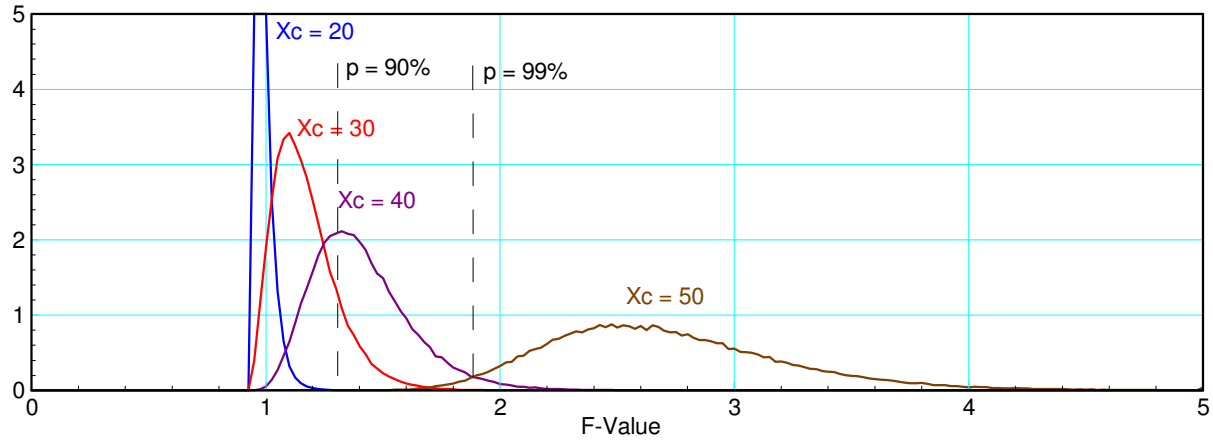
This can be found using StatTrek with

- 59 degrees of freedom in the numerator (N-1 = 59)
- 57 degrees of freedom in the denominator (N-k = 57)



Using Stat-Trek, the F statistic for 90% and 99% probability can be found

the resulting pdf for the F-value is as follows:



F-Value with Na = Nb = Nc = 20.

Note that

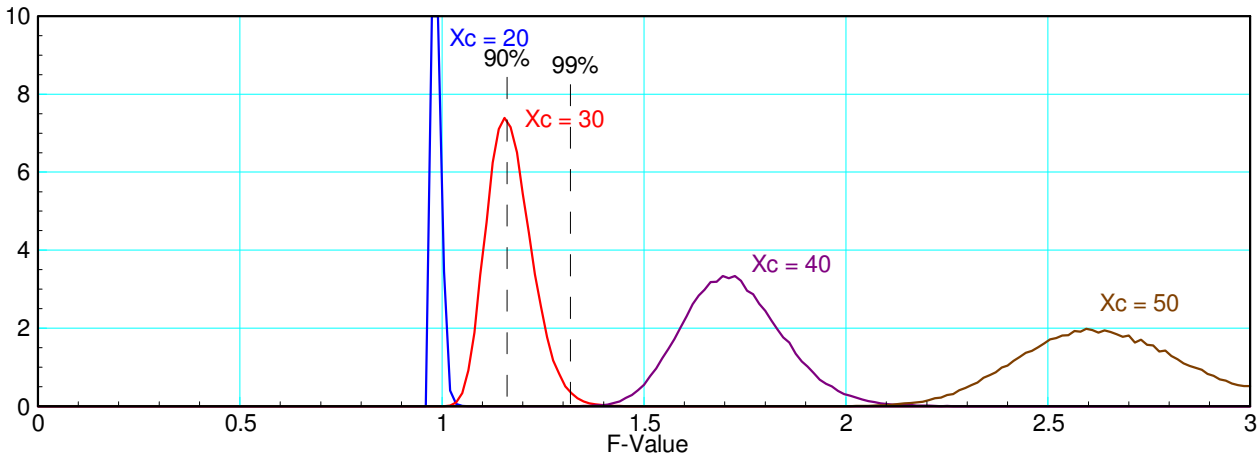
- At 90% certainty, you can usually detect a difference in the means when population C's mean is 2x the means of populations A and B
- At 99% certainty, you can almost always detect a difference in the means when C's means is 2.5x larger

These results chance as the variance changes.

As the sample size increases, you can detect smaller and smaller differences in the population's mean. When the sample size is 100 for A, B , and C, the critical F-statistic becomes

- F = 1.160 for p = 90%
- F = 1.310 for p = 99%

The pdf for the F-value with different means in population C is then



F-Value when the sample size = 100 for populations A, B, and C

Note that with more data, it is easier to detect the difference in the means:

- You can usually detect a difference of 10 in the means with 90% certainty,
- You can almost always detect a difference in means of 20 (Xc = 40) with 99% certainty

Matlab Code:

```
npt = 1e5;
N0 = 100;
Xmax = 3;
dx = Xmax / 200;
x = [0:dx:Xmax]';
y = zeros(length(x),1);
for i=1:npt
    A = 10*randn(N0,1) + 20;
    B = 11*randn(N0,1) + 20;
    C = 12*randn(N0,1) + 50;
    Na = length(A);
    Nb = length(B);
    Nc = length(C);
    N = Na + Nb + Nc;
    k = 3;
    G = mean([A; B; C]);
    MSSb = var([A; B; C]);
    MSSw = 1/(N-k) * ((Na-1)*var(A) + (Nb-1)*var(B) + (Nc-1)*var(C));
    F = MSSb / MSSw;
    n = round(F/dx);
    n = max(1, n);
    n = min(length(y),n);
    y(n) = y(n) + 1;
end
y = y / npt / dx;
```

```
plot(x,y);
toc
```

ANOVA Equations: Variation #2 (Standard Method)

While the previous way of computing MSSb and MSSw is the *correct* way (in my opinion), it's not the standard way of computing them. The standard way is as follows.

The previous analysis assumed you have access to the raw data for each population. Sometimes, this information is lost and all you have is each population's

- Mean,
- Variance, and
- Sample size.

Using only these terms, you can approximate MSSb and compute MSSw

MSSb: The 'correct' way to compute MSSb is

$$MSS_b = \left(\frac{1}{N-1}\right)\left(\Sigma\left(a_i - \overline{G}\right)^2 + \Sigma\left(b_i - \overline{G}\right)^2 + \Sigma\left(c_i - \overline{G}\right)^2\right)$$

with N-1 degrees of freedom. Assuming the variance of each population is zero, then

$$\Sigma\left(a_i - \overline{G}\right)^2 \approx n_a\left(\overline{A} - \overline{G}\right)^2$$

allowing you to rewrite MSSb as

$$MSS_b \approx \left(\frac{1}{k-1}\right)\left(n_a\left(\overline{A} - \overline{G}\right)^2 + n_b\left(\overline{B} - \overline{G}\right)^2 + n_c\left(\overline{C} - \overline{G}\right)^2\right)$$

with k-1 degrees of freedom. Note that the degrees of freedom drop since you replace N variables in the previous equation with just 3 (k) variables in the approximate equation.

MSSw: Whereas MSSb has to be approximated if you don't have access to the raw data, MSSw can be computed exactly:

$$MSS_w = \left(\frac{1}{N-k}\right)\left((n_a - 1)s_a^2 + (n_b - 1)s_b^2 + (n_c - 1)s_c^2\right)$$

F-Value: Once you have MSSb and MSSw, the F-value is the same as before

$$F = \frac{MSS_b}{MSS_w}$$

As noted before, this is actually the standard way doing an ANOVA computation - even when you have access to the actual data.

ANOVA Example #2:

Let's repeat the previous example with

- $A \sim N(20, 10^2)$
- $B \sim N(20, 11^2)$
- $C \sim N\left(\overline{C}, 12^2\right)$

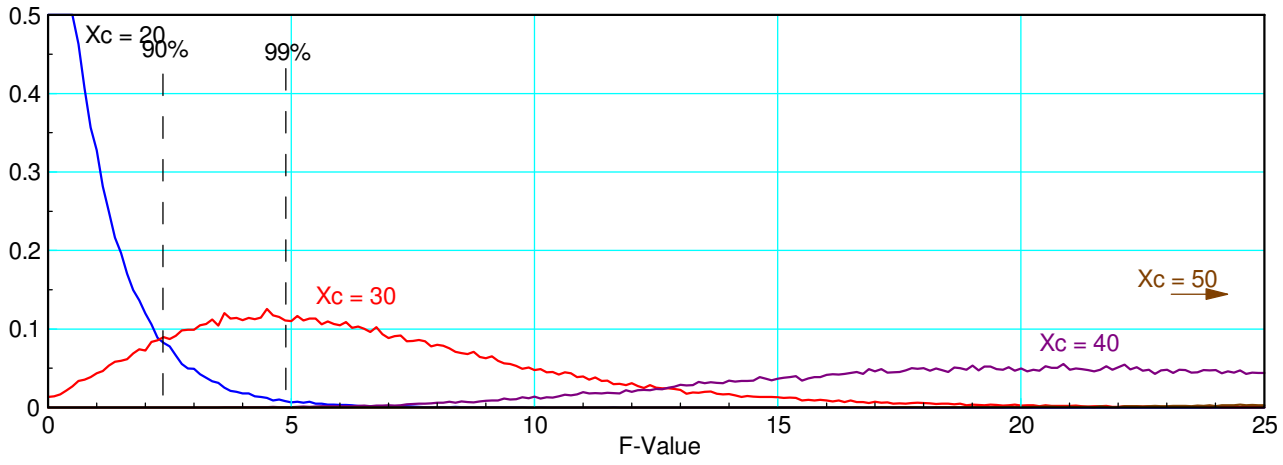
and the mean of C varies from 20 to 50. Using the approximate equations for MSSb, the critical F-value has

- A numerator with 2 degrees of freedom (k-1), and
- A denominator with 59 degrees of freedom (N-1)

This corresponds to

- F = 2.395 for p = 90% and
- F = 4.983 for p = 99%

Running a Monte-Carlo simulation with 100,000 random values for populations A, B, and C result in the following F-values when the mean of C varies:

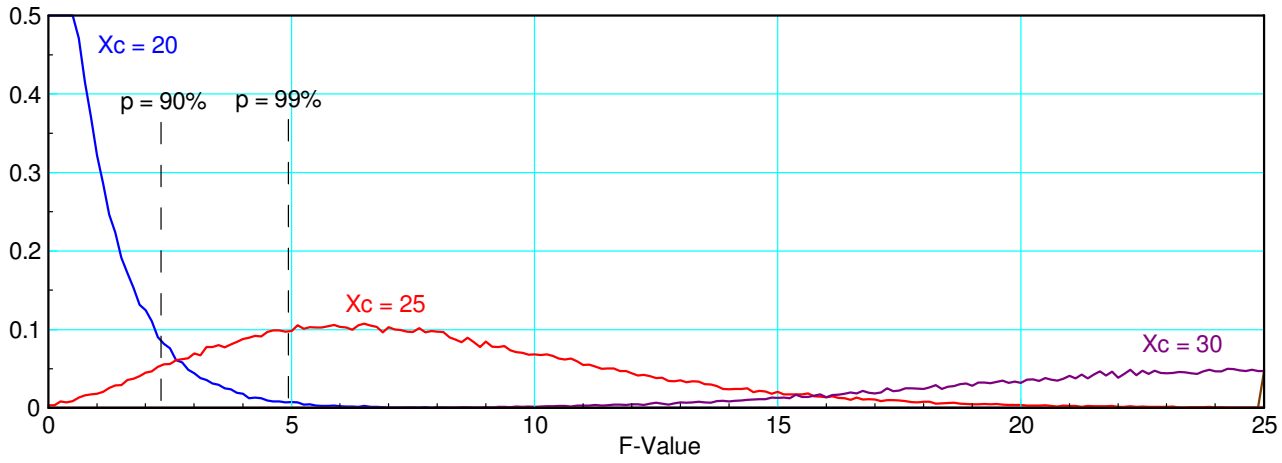


pdf for the F-value with the mean of population C is 20, 30, 40, and 50

Note that with this form of Mssb

- You can usually detect a 50% difference in the mean with 90% certainty,
- You can almost always detect a 100% difference in mean (Xc = 40) with 99% certainty, and
- There is a lot more noise in the resulting F value.

If you increase the sample size to 100 for each population, the pdf for the F-value as the mean of C varies is as follows:



pdf for the F-value when the mean of C is 20, 25, and 30

Note from this figure that

- You can usually detect a change in mean of 25% (Xc=25) with 90% certainty, and
- You can almost always detect a change in mean of 50% (Xc = 30) with 99% certainty.

The Matlab code for these Monte-Carlo simulations is as follows:

```
tic
npt = 1e5;
N0 = 100;
Xmax = 25;
dx = Xmax/200;
x = [0:dx:Xmax]';
y = zeros(length(x),1);
for i=1:npt
    A = 10*randn(N0,1) + 20;
    B = 11*randn(N0,1) + 20;
    C = 12*randn(N0,1) + 35;

    Xa = mean(A);
    Xb = mean(B);
    Xc = mean(C);
    Na = length(A);
    Nb = length(B);
    Nc = length(C);
    Va = var(A);
    Vb = var(B);
    Vc = var(C);

    N = Na + Nb + Nc;
    k = 3;
    G = 1/N * (Na*Xa + Nb*Xb + Nc*Xc);
    MSSb = (1/(k-1)) * (Na*(Xa-G)^2 + Nb*(Xb-G)^2 + Nc*(Xc-G)^2);
    MSSw = 1/(N-k) * ((Na-1)*Va + (Nb-1)*Vb + (Nc-1)*Vc);

    F = MSSb / MSSw;
    n = round(F/dx);
    n = max(n,1);
```

```
    n = min(length(y),n);
    y(n) = y(n) + 1;
end
z = length(y);
y(z) = y(z-1);
y = y / npt / dx;
plot(x,y);
toc
```

ANOVA Table

The typical (and equivalent) way to compute F is with an ANOVA table. This uses the latter method for MSSb.

A	B	C	$\left(a_i-\overline{A}\right)^2$	$\left(b_i-\overline{B}\right)^2$	$\left(c_i-\overline{C}\right)^2$
18.2501 20.9105 20.8671 19.9201 20.8985 20.1837 20.2908 20.1129	20.7599 20.2525 24.2810 18.3500 17.3186 18.3890 18.4600 19.4496	21.6631 21.5629 23.0827 22.7785 23.5025 25.5565 24.4461 19.4335	3.7215 0.5348 0.4732 0.0671 0.5174 0.0000 0.0125 0.0044	1.2151 0.3539 21.3761 1.7098 5.4708 1.6093 1.4342 0.0433	1.1884 1.4169 0.1086 0.0006 0.5614 7.8584 2.8658 11.0206
19.9649 mean(A)	19.6576 mean(B)	22.7532 mean(C)	5.33	33.21	25.02
20.7588 global mean (G)			63.5638 SSw		
8 na	8 nb	8 nc	3.0268 MSSw		
24 N					
43.95 SSb					
21.97 MSSb					

Step 1: Start with the data (shown in yellow)

Step 2: Calculate MSSb (shown in blue)

- Find the mean of A, B, C
  `mean(A)`
- Find the global mean, G
  `G = mean( [A;B;C] )`
- Find the number of data points in A, B, C
  `Na = length(A)`
- Find the total number of data points
  `N = Na + Nb + Nc`
- Compute the sum-squared total between columns
  `SSb = Na*(mean(A)-G)^2 + Nb*(mean(B)-G)^2 + Nc*(mean(C)-G)^2`
- Compute the mean sum-squared to tal between columns
  `MSSb = SSb / (k-1)`

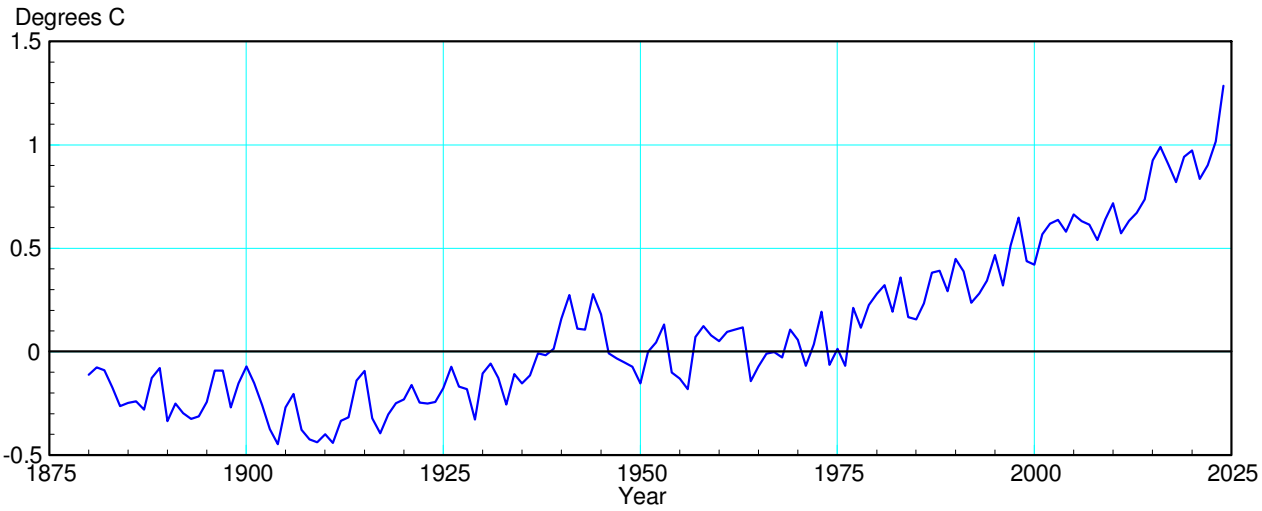
Step 3: Calculate MSSw (shown in pink)

- Compute $\left(a_i-\overline{A}\right)^2$
  `(A - mean(A)).^2`
- Find the total
  `sum( (A-mean(A)).^2 )`
- Add them up
  `SSw = sum((A-mean(A)).^2) + sum((B-mean(B)).^2) + sum((C-mean(C)).^2)`
- Find MSSw
  `MSSw = SSw / (N-k)`

ANOVA Examples

Global Temperatures: 1880s, vs. 1890s vs. 1900s

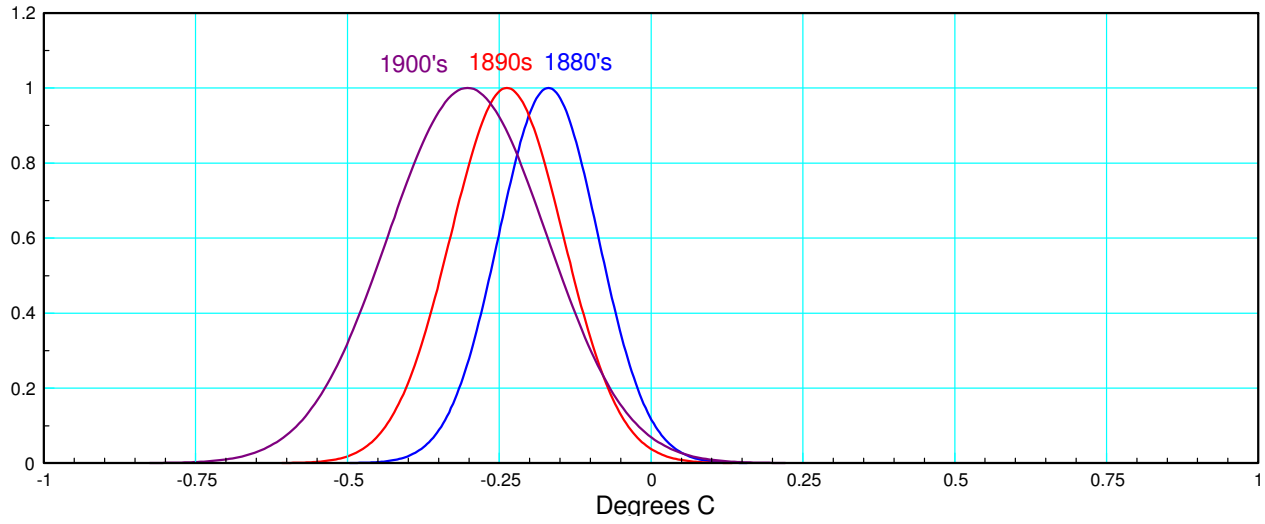
To illustrate the use of ANOVA, consider global temperatures as recorded by NASA Goddard:



Global Temperature Deviations (From NOAA)

Determine the probability that the average global temperature in the 1890s, 1900s, and 1910a were the same using ANOVA.

Step 1: Collect data. This can be obtained from NOAA (and Bison Academy). Find the mean, standard deviation, and sample size for each population



pdf for Global Temperature Deviations for the decades {1880s, 1890s, 1900s}

Step 2: Compute MSSb, MSSw and F. (code at the end of this section). The results are:

```
MSSb = 0.044398
num dof = 2
MSSw = 0.010797
den dof = 27
F-Value = 4.1122
```

Step 3: Convert the F-value to a probability. Using StatTrek, p = 0.972

It is 97.2% certain that the average temperature over these three decades were different.

If you use the actual data to compute the probability (more accurate but not standard way of doing it)

```
N = 30
MSSb = 0.013114
num dof = 29
MSSw = 0.010797
den dof = 27
F-Value = 1.2146
```

This corresponds to a probability of 69.3%

It is 69.3% likely that the three decades do not have the same mean

Code: Standard Method (method #2)

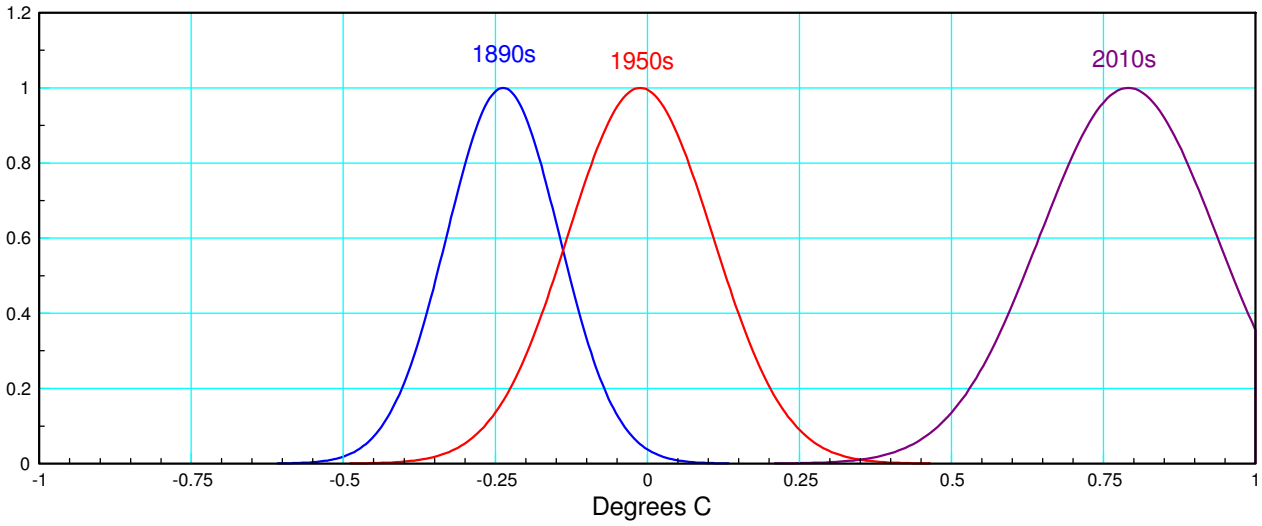
```
A = dT(1:10);
B = dT(11:20);
C = dT(21:30);

Xa = mean(A);
Xb = mean(B);
Xc = mean(C);
Na = length(A);
Nb = length(B);
Nc = length(C);
Va = var(A);
Vb = var(B);
Vc = var(C);

N = Na + Nb + Nc;
k = 3;
G = 1/N * (Na*Xa + Nb*Xb + Nc*Xc);
MSSb = (1/(k-1)) * (Na*(Xa-G)^2 + Nb*(Xb-G)^2 + Nc*(Xc-G)^2);
MSSw = 1/(N-k) * ((Na-1)*Va + (Nb-1)*Vb + (Nc-1)*Vc);

F = MSSb / MSSw;

disp(['N = ',num2str(N)])
```

```
disp(['MSSb = ',num2str(MSSb)])
disp(['num dof = ',num2str(k-1)])
disp(['MSSw = ',num2str(MSSw)])
disp(['den dof = ',num2str(N-k)])
disp(['F-Value = ',num2str(F)]);
```

Code: (nonstandard method - method #1)

```
A = dT(1:10);
B = dT(11:20);
C = dT(21:30);

Xa = mean(A);
Xb = mean(B);
Xc = mean(C);
Na = length(A);
Nb = length(B);
Nc = length(C);
Va = var(A);
Vb = var(B);
Vc = var(C);

N = Na + Nb + Nc;
k = 3;
G = 1/N * (Na*Xa + Nb*Xb + Nc*Xc);
MSSb = var([A;B;C]);
MSSw = 1/(N-k) * ((Na-1)*Va + (Nb-1)*Vb + (Nc-1)*Vc);

F = MSSb / MSSw;

disp(['N = ',num2str(N)])
disp(['MSSb = ',num2str(MSSb)])
disp(['num dof = ',num2str(N-1)])
disp(['MSSw = ',num2str(MSSw)])
disp(['den dof = ',num2str(N-k)])
disp(['F-Value = ',num2str(F)]);
```

Global Temperatures: 1890s, vs. 1950s vs. 2010s



pdf for global temperature deviations over the decades of {1890s, 1950s, 2010s}

Repeating the previous analysis, using the standard method (method #2)

```
N = 30
MSSb = 2.9207
num dof = 2
MSSw = 0.014642
den dof = 27
F-Value = 199.4749
```

This corresponds to a probability of 1.000.

Using the raw data (non-standard method)

```
N = 30
MSSb = 0.21506
num dof = 29
MSSw = 0.014642
den dof = 27
F-Value = 14.6879
```

This also corresponds to a probability of 1.000

I am almost 100% certain that these decades do not have the same mean temperature

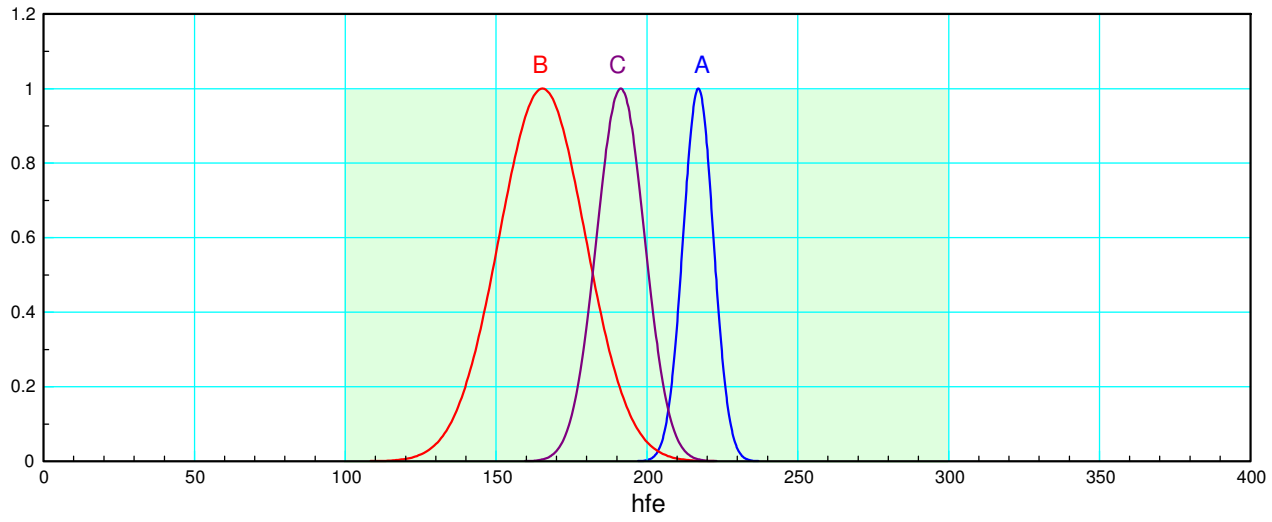
3904 Transistors

Three shipments of 3904 NPN transistors were received. Statistics for the gain (hfe) for these shipments are as follows:

	Mean	St Dev	n
Shipment #1 (A)	217.086	4.980	70.000
Shipment #2 (B)	165.333	14.247	12.000
Shipment #3 (C)	191.263	7.931	38.000

Determine if these shipments have the same mean.

Comment: The data sheets specify the current gain to be in the range of (100,300). All batches are well within this range and meet specs. What the ANOVA test tells you is whether these transistors have a common source (same manufacturer, same production run, etc.)



pdf for the three shipments of 3904 NPN transistors and the specs: 100 < hfe < 300

Since the raw data is not available, we'll have to use the standard (second) method to do ANOVA.

```
N = Na + Nb + Nc;
k = 3;
G = 1/N * (Na*Xa + Nb*Xb + Nc*Xc);
MSSb = (1/(k-1)) * (Na*(Xa-G)^2 + Nb*(Xb-G)^2 + Nc*(Xc-G)^2);
MSSw = 1/(N-k) * ((Na-1)*Va + (Nb-1)*Vb + (Nc-1)*Vc);

F = MSSb / MSSw;
```

This results in

```
N = 120
MSSb = 18041.9729
num dof = 2
MSSw = 53.6027
den dof = 117
F-Value = 336.5868
```

From StatTrek, this corresponds to a probability of 1.000

I'm almost 100% certain that these transistors do not come from the same source (means are different).

3904 Transistors (Shipment #1)

Take shipment #1 and split into three piles. Do ANOVA to see if these are from different sources

	Mean	St Dev	n
Shipment #1a (A)	216.172	5.465	29
Shipment #1b (B)	217.150	3.977	20
Shipment #1c (C)	218.286	5.100	21

Again, without access to the raw data, the second (standard) way of doing ANOVA is necessary. This results in

```
N = 70
MSSb = 27.256
num dof = 2
MSSw = 24.7309
den dof = 67
F-Value = 1.1021
```

From StatTrek:

```
p = 0.662
```

I am 66.2% certain that the three groups of transisors in shipment #1 have different means (meaning they came from different manufacturers, different production runs, etc.) 66.2% certainty means there's no evidence to say the means are different.

Summary

Analysis of Variance (ANOVA) is a tool you can use if you want to compare means for more than two populations. This results in an F-test where a large F-value indicates you can reject the null hypothesis (that the means are the same). If you want to determine which means are the outliers, a different test needs to be used.